AI/ML Task 2 – Feature Engineering, Model Optimization & Performance Comparison

1. Introduction

This project focuses on implementing machine learning models to predict house prices using the California Housing Dataset. The main objective of the task is to perform feature engineering, apply data preprocessing techniques, train multiple regression models, and compare their performance using evaluation metrics. Feature engineering and model optimization play an important role in improving model accuracy and performance.

Machine learning models are widely used in real-world applications such as price prediction, recommendation systems, and data analysis. This project demonstrates the complete machine learning workflow including data preparation, feature scaling, model training, and performance evaluation.

2. Dataset Description

The California Housing Dataset was used for this project. The dataset contains various features related to housing information such as median income, house age, number of rooms, population, and geographical location. The target variable of the dataset is the median house price.

The dataset includes the following features:

- Median Income (MedInc)
- House Age (HouseAge)
- Average Rooms (AveRooms)
- Average Bedrooms (AveBedrms)
- Population
- Average Occupancy (AveOccup)
- Latitude
- Longitude

The goal of the model is to predict the house price based on these input features.

3. Methodology

The project was implemented using Python and machine learning libraries such as pandas, NumPy, scikit-learn, and matplotlib in a Jupyter Notebook environment.

The following steps were performed:

Data-Loading:
The dataset was loaded using the scikit-learn library and converted into a structured data format.

Feature-Separation:
Input features and target variables were separated for training the model.

Feature-Scaling:
StandardScaler was applied to normalize the feature values and improve model performance.

Train-TestSplit:
The dataset was divided into training data (80%) and testing data (20%) to evaluate model performance.

ModelTraining:
Three machine learning models were trained:

- Linear Regression
- Ridge Regression
- Decision Tree Regressor

ModelEvaluation:
The models were evaluated using Root Mean Squared Error (RMSE) and R² score.

4. Model Performance Comparison

The performance of the models was compared using evaluation metrics.

Model	**RMSE**	**R² Score**
Linear Regression	0.745581	0.575788
Ridge Regression	0.745554	0.575819
Decision Tree	0.724234	0.599732

From the results, the Decision Tree Regressor achieved the lowest RMSE and highest R² score, indicating better prediction performance compared to other models.

5. Results and Analysis

The Decision Tree model provided the best performance among the three models. Feature scaling improved the learning efficiency of the models. Model comparison helped identify the most suitable model for house price prediction.

The project demonstrates how feature engineering and model optimization techniques can improve machine learning performance.

6. Conclusion

This project successfully implemented feature engineering, model optimization, and performance comparison techniques using machine learning models. Multiple regression models were trained and evaluated using standard metrics. The Decision Tree model performed best for the given dataset.

The project provides practical knowledge of data preprocessing, model training, and evaluation in machine learning applications.

7. Tools Used

- Python
- pandas
- NumPy
- scikit-learn
- matplotlib
- Jupyter Notebook / Google Colab